MASHing Metadata: Legacy Issues in OAI Harvesting From Three Digital Libraries

Michael Khoo
The iSchool
Drexel University
khoo@drexel.edu

Doug Tudhope
Faculty of Advanced Technology
The University of Glamorgan
dstudhope@glam.ac.uk

Ceri Binding
Faculty of Advanced Technology
The University of Glamorgan
cbinding@glam.ac.uk

Hilary Jones
Mimas
University of Manchester
Hilary.Jones@manchester.ac.uk

Ivan Orrego
The iSchool
Drexel University
ifo23@drexel.edu

Jae-wook Ahn
The iSchool
Drexel University
ja626}@drexel.edu

Abstract

This Note reports on efforts to build a generalizable OAI-PMH workflow to retrieve metadata sets from unrelated digital libraries. This effort is part of a wider effort to build a database to aggregate metadata from different digital libraries, which can then be used as the basis for content analysis and data mining experiments with the metadata records. A pilot metadata harvest from three digital libraries using OAI- PMH encountered a number of issues, arising from idiosyncratic legacy characteristics of each of the three metadata sets. In the end, the harvests had to be manually tailored to each library. OAI-PMH proved to be a useful approach, but only after communication with each digital library had identified important characteristics of each metadata set, including many legacy characteristics, which had to be acounted for in the harvest.

Keywords: digital libraries, Dublin Core, legacy issues, metadata harvesting, OAI-PMH

Introduction

Interoperability is both a desirable and also an elusive goal for digital libraries (Paepcke et al., 1998; Gradmann, 2009). Individual digital libraries contain collections of high-quality resources, and combinations of resources from separate libraries can yield rich educational and research insights, but at the same time, it is not easy to search across different digital libraries. Federated search or browsing services are often simply not available, and where they do exist, they can, for various reasons, be of limited utility. While there is thus enormous potential present in the large numbers of digital libraries created so far, very often this potential is 'locked up' in what might be thought of as individual library silos. There are number of solutions to the 'un-siloing' of these digital libraries, such as mapping and crosswalking metadata in different libraries to a common format which can then be stored in and queried from a central repository. In reality, these solutions can be complex resource intensive endeavors, which can produce mixed results (Khoo & Hall, 2010; Lagoze et al., 2006). It is therefore worth exploring further strategies for digital library interoperability.

The work described in this Note is being carried out by three teams in the U.S.A. and the U.K., and is exploring methods for increasing discovery across unaffiliated digital libraries without the use of metadata crosswalks (Digging into Metadata, 2011: http://research.cis.drexel.edu/digging/). The three digital libraries in the study are the National Science Digital Library (U.S.A.: http://nsdl.org/) (also including the Digital Library for Earth Systems Education, DLESE: http://www.dlese.org/); the Internet Public Library (USA: http://www.ipl.org/) (also including the Librarians' Internet Index (LII); and Intute (U.K.: http://www.intute.ac.uk/).

Acknowledgements: This research is funded by IMLS under grant # LG-00-12-0457-12, as part of the *Digging Into Data Challenge.*
Khoo, M., Tudhope, D., Binding, C., Jones, H., Orrego, I., & Ahn, J-w. (2013). MASHing metadata: Legacy issues in OAI harvesting from three digital libraries. *iConference 2013 Proceedings* (pp. 497-501). doi:10.9776/13263

Harvesting Metadata Records for Content Analysis

A central premise of the work is that human-generated metadata records contain information which can be aggregated and subject to machine content analysis. In this study, a content analysis of metadata records collected from three digital libraries is being used to support the generation of Dewey Decimal Classification 'tags,' which will then be added back to each metadata record, in order to enhance search and browse functionality across the three libraries.

Not all fields in a metadata record are equally useful for content analysis, and so the initial work has focused on selecting and analyzing the title, description, and keyword and subject fields, in a small sample set of 50 metadata records obtained from these three digital libraries. An outline of the analysis is provided in Table 1, which shows a hypothetical metadata record (column 1), the fields selected for
further analysis (the Title, Description, and various Subject fields) (column 2), and the 'cleaned' content of these fields which will then be used for the content analysis (column 3).

Table 1
Metadata cleaning of a Dublin Core for a hypothetical web site ("astrophysics.org")

Metadata obtained via OAI-PMH	Selected fields	Cleaned fields
<header>Administrative metadata</header> <metadata> <dc:identifier>http://www.astrophysics.org/ </dc:identifier> <dc:title>Astrophysics</dc:title> <dc:description>A review of space science and astrophysics </dc:description> <dc:format></dc:format> <dc:type></dc:type> <dc:publisher></dc:publisher> <dc:subject>astrophysics</dc:subject> <dc:subject>space</dc:subject> <dc:subject>science<dc:subject> </metadata>	<dc:title>Astrophysics</dc:title> <dc:description>A review of space science and astrophysics </dc:description> <dc:subject>astrophysics</dc:subject> <dc:subject>space</dc:subject> <dc:subject>science<dc:subject>	Astrophysics A review of space science and astrophysics astrophysics space science

Much of the work so far has focused on the development of a prototype tool – 'Metadata Aggregation, Storage, and Handling' (MASH) – that will aggregate and process the metadata from each of the digital libraries, before handing off results to a second tool (Document Indexing and Semantic Tagging Interface for Libraries: DISTIL) for the generation of DDC tags (Khoo et al., 2012). An overview of the MASH tool, identifying the initial points of data 'hand-off' between the different partners, is provided in Figure 1.

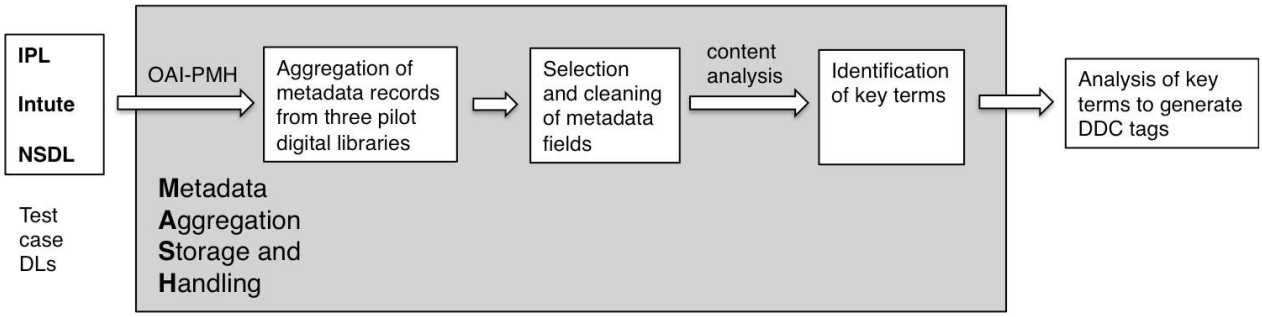

Figure 1. Overall workflow for the MASH tool.

Initial work with MASH used OAI-PMH to collect a small number of Dublin Core metadata records from each of the three 'test case' digital libraries; selecting the title, description and subject fields from each record in each collection; removing XML markup; and applying several content analysis methods to

different configurations of the cleaned metadata, in order to identify best methods for extracting sets of key terms for use downstream in the project. While it was thought that some of aspects of this work would be more complex than others, in practice, even apparently simple stages have had unexpected dimensions, including the initial metadata harvesting.

Legacy Issues Affecting Metadata Harvesting

The original project proposal envisaged developing a prototype harvesting workflow with the three pilot digital libraries, with a successful proof-of-concept paving the way for the development of a more generalizable OAI-PMH workflow, allowing the project to scale up to additional digital libraries. The MASH team began the work by carrying out several manual harvests of small metadata sets from each of the three libraries. A number of issues were encountered in this initial harvesting, including non OAI- accessible metadata, variations in subject fields, and undocumented metadata aggregation and 'normalization' in union repositories. The resolution of these issues required discussions with each individual digital library before the metadata itself could be harvested in a form that was useful for the project.

The first example is the IPL. This was founded in 1995 as an online reference service, and then began developing digital collections (Janes, 1998). Beginning in 2008, the IPL merged collections with the Librarians' Internet Index (LII), and the IPL and LII metadata was crosswalked to Dublin Core and added to a Fedora database (Khoo & Hall, 2010). Three of the fields of interest to the project – dc:title, dc:description, and dc:subject – were placed in the main DC datastream. However, two further potentially useful fields from the original metadata were placed in other datastreams. First, the ipl:subject field, which included custom-formatted IPL browsing metadata inherited from the legacy SQL metadata, was placed in a separate IPL datastream (this IPL datastream also archived all the original IPL metadata that could not be mapped to the 15-element Dublin Core set). Second, the somewhat complex relationships between item-level resources and various types of collections in IPL were mapped to an RDF triple, and placed in a third datastream (it was decided at the time not to use the dc: relation element, as the IPL did not have a consistent definition of the relationship between a collection and an item). There was therefore metadata useful for MASH that was neither visible nor retrievable in a standard OAI query to the Dublin Core datastream. Understanding the location and nature of this additional metadata required some familiarity with IPL history, in order to configure the harvest.

The second example is Intute, which was developed by a grass-roots community dedicated to online educational resource discovery (Joyce, 2008; Williams, 2006). Much Intute metadata was collected by previous partners and consortiums. Partly as a consequence, each Intute resource has both a DC record and also additional subject classification metadata stored in separate SQL tables, which can be drawn on as needed. These SQL tables are partly a legacy of the prior projects that were subsequently migrated to Intute, which utilized specific subject catalogs to suit the needs of particular audiences, for instance by offering domain-specific keywords. Once again, the harvesting issue here was that there was potentially useful metadata which were neither visible nor retrievable via standard OAI queries, and which required communication with the Intute cataloging staff, this time by email and teleconference, to locate and understand.

The third example is the NSDL, a federated multi-disciplinary STEM library, with a central metadata repository at nsdl.org. The central repository integrates metadata from the individual domain- specific portals, or 'Pathways' (e.g. Zia, 2004; Bikson et al., 2011), with Pathways' metadata being passed via OAI-PMH. (The metadata harvested by MASH from NSDL has therefore been through at least two OAI-PMH pipelines – from the Pathway to NSDL, and from NSDL to the current project.) As NSDL Pathways are independent entities, the same resource can be cataloged in different ways by different Pathways. In these cases, the record displayed to users at nsdl.org is a 'normalized' version of all records created by individual Pathways, in which all the subject terms from each Pathway record were preserved (with editing for redundancy) in the 'normalized' record. This resulted in multiple (often similar) subject terms being displayed in the record, which could skew the results of the content analysis in MASH. Email communication with the NSDL's metadata staff was required to clarify this situation, and to adjust the harvesting process for the current project.

Discussion

In each of the three cases just described, the various historical contexts within which each library created metadata had resulted in a number of legacy metadata issues. These legacy issues often came to light only by accident, as happened for example when manually comparing an XML record obtained via OAI-PMH from a particular library, with the Web display of the same record at the library's Web site, and noticing discrepancies between the two. A useful finding from the initial harvesting experiments was therefore that it was not possible to create a single generic OAI-PMH query to retrieve all the required metadata from all the libraries. Additional metadata had to be accessed via refined queries, or sent as .sql or .csv files. This required additional communication with the metadata owners in order to understand how the metadata had been created and structured in the first place, before any action could proceed. On the one hand, therefore, while each library had contributed its metadata for the pilot MASH analyses in good faith, it was not obvious to the metadata owners exactly what metadata the MASH team were interested in; on the other hand, though, the MASH team did not have a full picture of what metadata was available. The team therefore had to engage in boundary spanning (e.g. Brown & Duguid, 2001), and perspective making and perspective taking (Boland & Tenkasi, 1995), between the needs of the MASH tool, and the specific characteristics of the metadata in each library (Table 2). This analysis and communication was time consuming.

Table 2
Examples of metadata issues that required communication with individual digital libraries

Digital library	Metadata field	Issues	Communication
IPL	ipl:subject	Legacy browsing data from previous version of IPL - not available in main Fedora 'DC' datastream - archived in 'IPL' datastream'	Email queries, face-to- face clarification
	isMemberOf	RDF triple describing item-collection relationship - not available in main Fedora 'DC' datastream - archived in 'NNN' datastream'	
Intute	classification	Domain specific classification keywords stored in a separate SQL table	Email queries, teleconference
NSDL	dc:title dc:description dc:subject	In cases where multiple NSDL Pathways cataloged the same resource, the contents of metadata fields have been aggregated and 'normalized' into a representative record; the underlying Pathway metadata is still available but requires specialized OAI-PMH queries.	Email queries

In the end, all the required metadata was harvested from each of the three libraries, and successfully added to the MASH database. One of the original assumptions of the project proposal – that it would be possible to create a generalizable OAI-PMH workflow to support the scaling of the project to other digital libraries – was not however supported by the pilot study. In the case of MASH at least, there was always a need to communicate with a library before a harvest was implemented. Nevertheless, while the technical workflow was not generalizable, other generalizations can be made from the three cases, especially with regard to the need for detailed organizational communication to obtain full descriptions of metadata sets before harvesting is implemented.

Conclusion

This Note has described attempts to use OAI-PMH to retrieve and aggregate metadata from three digital libraries. While OAI-PMH worked as a harvesting technology, variations in the metadata formats of the individual libraries required manual analysis and resolution before a satisfactory OAI query format was reached. In addition, some metadata useful for the MASH project was not available even through OAI-PMH. The roots of these issues were traceable to a variety of legacy factors that had shaped how the metadata in each library had originally been created, formatted, and stored. These issues then had to be clarified through communication with each of the digital libraries concerned. The workflow was thus more time-consuming than originally anticipated, although it was ultimately successful. These findings have useful implications for other projects seeking to harvest metadata from disparate digital libraries.

References

Bikson, T., Kalra, N., Galway, L., & Agnew, G. (2011). Steps Toward a Formative Evaluation of NSDL. RAND Technical Report. http://www.rand.org/content/dam/rand/pubs/technical_reports/2011/RAND_TR998.pdf

Boland, R. J., & Tenkasi, R. V. (1995). Perspective Making and Perspective Taking in Communities of Knowing. Organization Science, 6 (4), 350-372.

Brown, J. S., & Duguid, P. (2001). Perspective. Knowledge and Organization: A Social-Practice Perspective. Organization Science 12 (2), 198-213.

Digging Into Metadata (2011). http://research.cis.drexel.edu/digging/

Ginger, K., & Goger, L. (2011). Evaluating the National Science Digital Library for Learning Application Readiness. Annual Meeting of the American Society for Information Science and Technology, October 9-13, 2011, New Orleans, LA.

Gradmann, S. (2009). Interoperability Challenges in Digital Libraries. First DL.org Workshop on Digital Library Interoperability, Best Practices and Modelling Foundations (ECDL 2009). http://www.dlorg.eu/uploads/Workshop%20Corfu/Interoperability%20Challenges%20in%20Digital%20Libraries_Gradmann.pdf

Janes, J. (1998). The Internet Public Library: An Intellectual History. Library Hi Tech, 16(2), 55-68. Joyce, A., Wickham, J., Cross, P., & Stephens, C. (2008). Intute integration. Ariadne 55. http://www.ariadne.ac.uk/issue55/joyce-et-al

Khoo, M., & Hall, C. (2010). Merging Metadata: A Sociotechnical Study of Crosswalking and Interoperability. 10th ACM/IEEE Joint Conference on Digital Libraries, Brisbane, Australia, June 21-25, 2010, pp. 361-36.

Khoo, M., Tudhope, D., Binding, C., Abels, E., Lin, X., & Massam, D. (2012). 'Towards Digital Repository Interoperability: The Document Indexing and Semantic Tagging Interface for Libraries (DISTIL). Theory and Practice of Digital Libraries (TPDL) 2012, Paphos, Cyprus, September 23-27, 2012.

Lagoze, C., & Van de Sompel, H. (2001). The Open Archives Initiative: Building a Low-Barrier Interoperability Framework. JCDL 2001, pp. 54-62.

Lagoze, C., Krafft, D., Cornell, T., Dushay, N., Eckstron, D., & Saylor, J. (2006). Metadata aggregation and automated digital libraries: A retrospective on the NSDL experience. Joint Conference on Digital Libraries (JCDL 2006), Chapel Hill, NC, pp. 230-239.

Paepcke, A., Chang, C., García-Molina, H., & Winograd, T. (1998). Interoperability for Digital Libraries Worldwide. Communications of the ACM 41(4), 33-43. Williams, C. (2006). Intute: The New Best of the Web. Ariadne 48. http://www.ariadne.ac.uk/issue48/williams

Zia, L. (2005). The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program. D-Lib Magazine 11(3). http://www.dlib.org/dlib/march05/zia/03zia.html